Predicting Psychological Distress

JARED HAMMERNIK

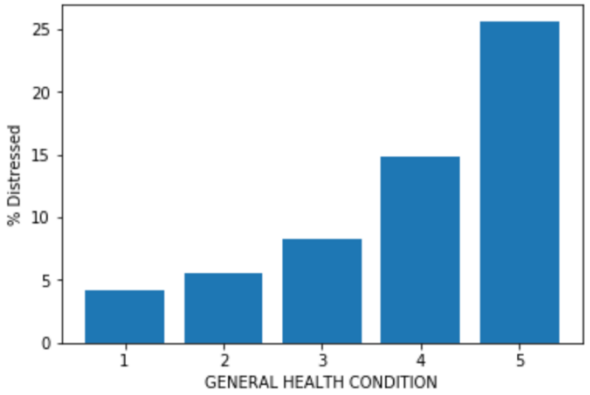
Introduction

- Data from adult 2017 California Health Information Survey (https://healthpolicy.ucla.edu/chis/Pages/default.aspx)
- **21,153** respondents, **483** columns
- Psychological Distress defined by Kessler's K6 scale (0-24)
- Targeting column "Likely Has Had Psychological Distress in the Last Year" which is defined by a Kessler score > 13
- **8.53%** of respondents likely to be distressed in last year

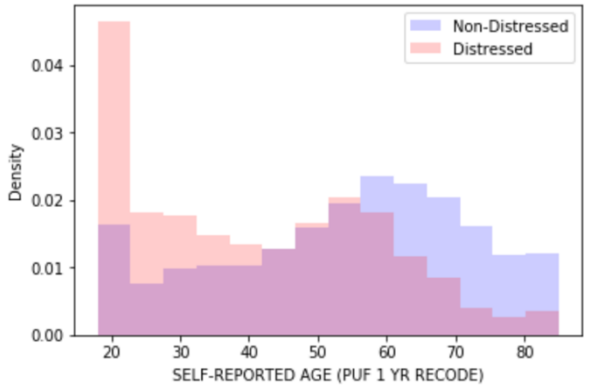
Data Cleaning

- In order to reduce dimensionality, I only selected columns that were answered by everyone in the survey. This left no missing values
- Reduced dimensionality further by manually selecting columns of interest and removing questions that provided duplicate information. Left with 78 columns
- One-Hot-Encoded categorical columns, increased to 121 columns
- Ordinal categorical columns were already ordinally encoded, made a few minor special changes but otherwise left untouched
- Removed collinear variables, reduced to **92** columns

(1 Excellent – 5 Poor)

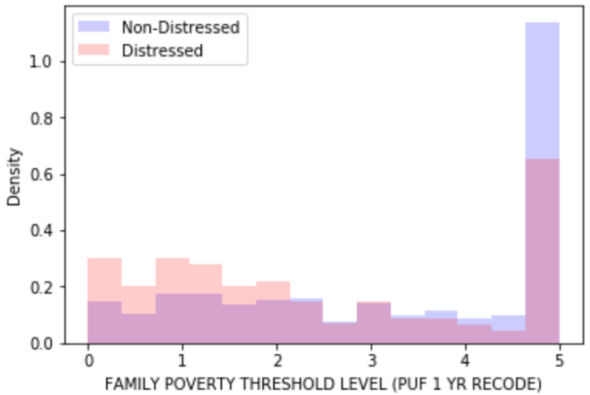


Chi-Squared P-value:
6.24e-157



Ind. T-Test P-value:
5.96e-159

(X Times Fed. Poverty Level)



Ind. T-Test P-value:
3.05e-79

Exploratory/Statistical Analysis

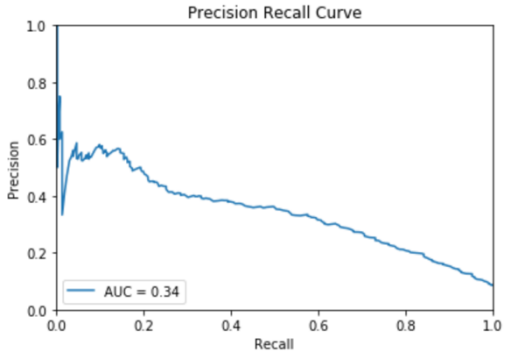
(Key Features)

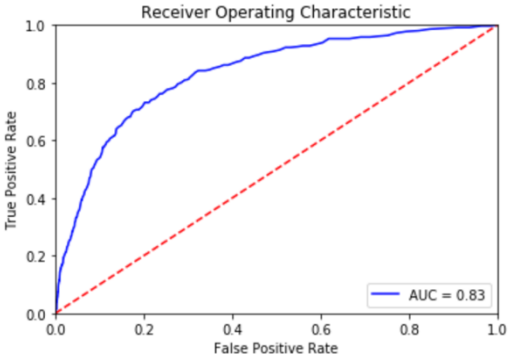
Model Pre-Processing

- Split data into **80%** training and **20%** test, stratified over target column
- Standardized Data (mean = 0, STD = 1)
- Since data set imbalanced, chose to optimize for Precision-Recall AUC
- Hyper-tuned parameters of each model using GridSearchCV

Random Forest

- Trees: 500
- Max Features: 10
- Criterion: Entropy
- Class Weight: None
- PR AUC: **0.34**
- ROC AUC: **0.83**

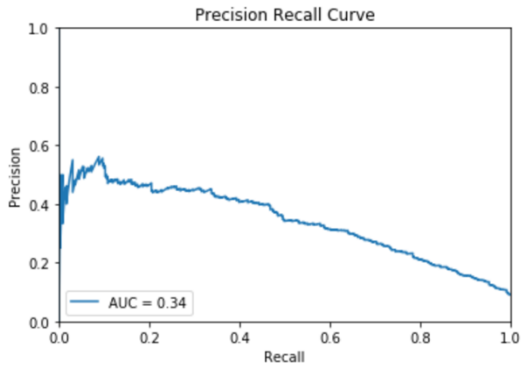


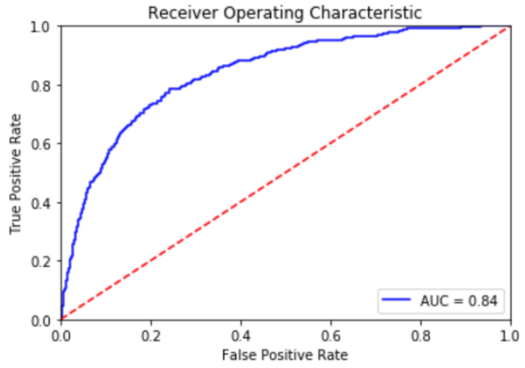


	feature	importance
40	SELF-REPORTED AGE (PUF 1 YR RECODE)	0.051416
52	WEIGHT: LBS (PUF RECODE)	0.042661
38	FAMILY POVERTY THRESHOLD LEVEL (PUF 1 YR RECODE)	0.040351
20	# OF TIMES SAW MD IN PAST 12 MOS	0.038542
0	GENERAL HEALTH CONDITION	0.035866
51	LENGTH OF TIME LIVED AT CURRENT ADDRESS (MONTH...	0.033229
13	# OF TIMES ATE VEGETABLES IN PAST MO	0.032576
10	# TIMES ATE FRUIT IN PAST MO	0.032317
12	# TIMES ATE COOKED DRIED BEANS IN PAST MONTH	0.031101
6	# OF TIMES DRANK WATER YESTERDAY	0.030451

Logistic Regression

- C: 0.01
- Penalty: L1 (Lasso)
- Class Weight: Balanced
- PR AUC: **0.34**
- ROC AUC: **0.84**

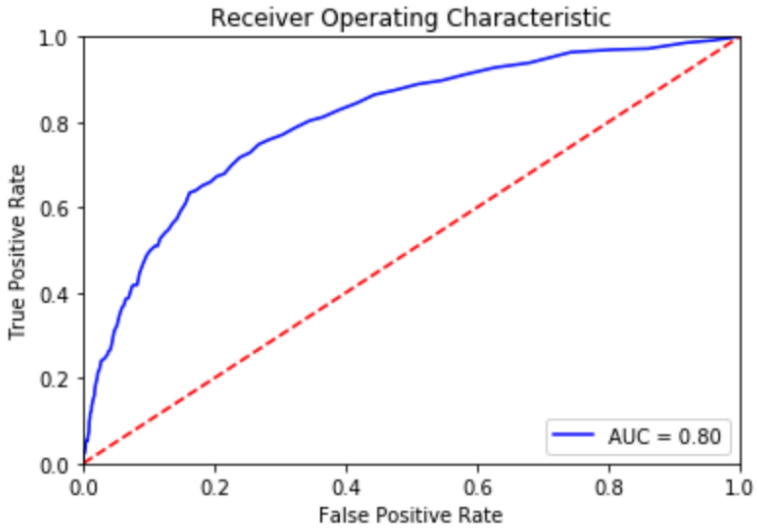


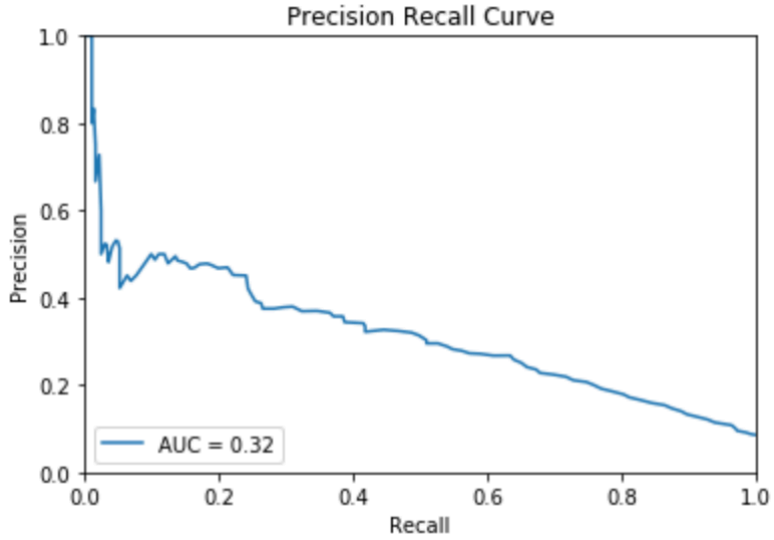


	feature	importance
40	SELF-REPORTED AGE (PUF 1 YR RECODE)	0.661732
0	GENERAL HEALTH CONDITION	0.374618
18	DELAY/NOT GET OTHER MEDICAL CARE IN PAST 12 MOS	0.230001
49	SELF-REPORTED GENDER	0.160739
72	MARITAL STATUS- 4 CATEGORIES_1.0	0.146461
3	EVER TRIED MARIJUANA OR HASHISH	0.142452
28	HOW OFTEN FEEL SAFE IN NEIGHBORHOOD	0.125172
7	EVER SMOKED ELECTRONIC CIGARETTES	0.122219
21	EVER HAD PROBLEMS PAYING FOR SELF OR HOUSEHOLD...	0.109163
73	MARITAL STATUS- 4 CATEGORIES_2.0	0.105924

KNN

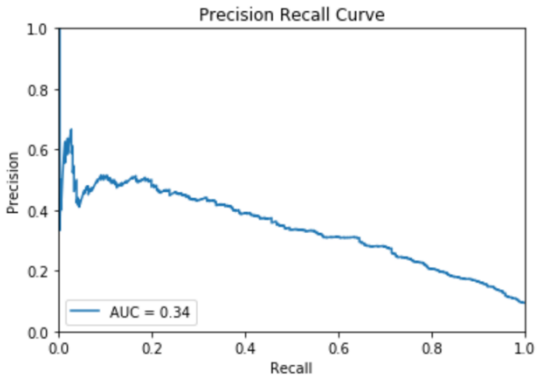
- Neighbors: 500
- PR AUC: **0.32**
- ROC AUC: **0.80**

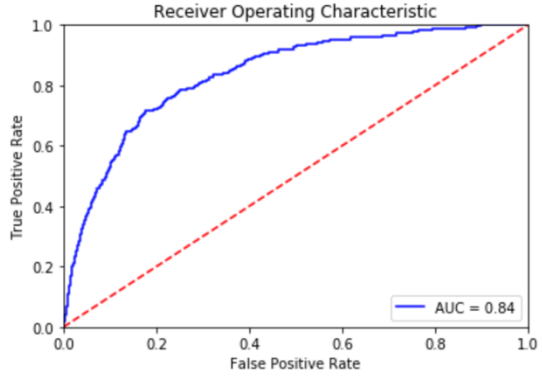




Gradient Boosting

- Trees: 100
- Max Features: 10
- Learning Rate: 0.1
- Loss: Deviance
- PR AUC: **0.34**
- ROC AUC: **0.84**

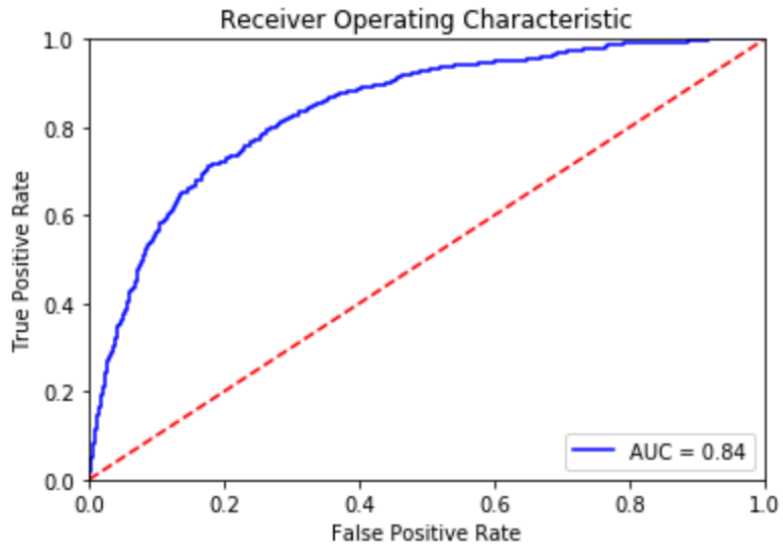


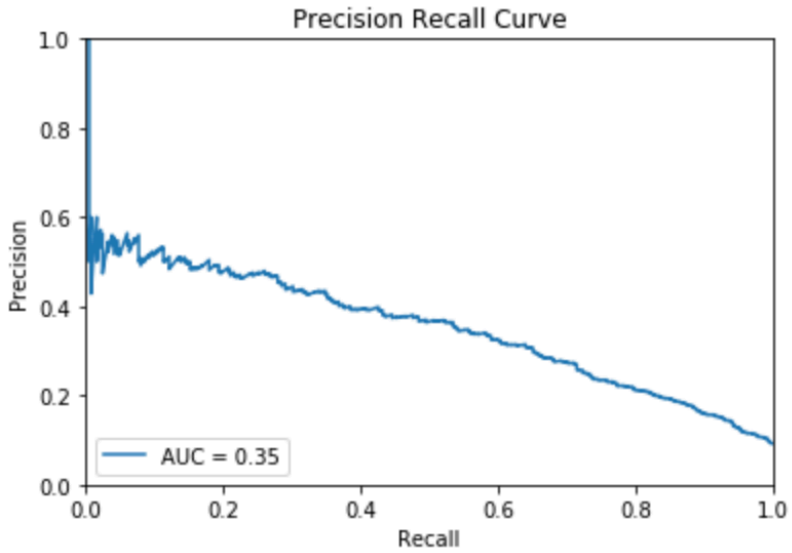


	feature	importance
0	GENERAL HEALTH CONDITION	0.123560
40	SELF-REPORTED AGE (PUF 1 YR RECODE)	0.115419
18	DELAY/NOT GET OTHER MEDICAL CARE IN PAST 12 MOS	0.097652
7	EVER SMOKED ELECTRONIC CIGARETTES	0.094165
21	EVER HAD PROBLEMS PAYING FOR SELF OR HOUSEHOLD...	0.042171
20	# OF TIMES SAW MD IN PAST 12 MOS	0.036883
79	SELF-REPORTED HOUSEHOLD TENURE (HH)_2.0	0.033323
38	FAMILY POVERTY THRESHOLD LEVEL (PUF 1 YR RECODE)	0.033204
17	DELAY/NOT GET PRESCRIPTION IN PAST 12 MO	0.029938
74	MARITAL STATUS- 4 CATEGORIES_4.0	0.029573

Voting Classifier

- Voting: Soft
- PR AUC: **0.35**
- ROC AUC: **0.84**





Conclusions

- Generally large difference between PR AUC and ROC AUC due to imbalance of classes
- Gradient Boosting and Logistic Regression were the best solo models with a PR AUC of **0.34** and an ROC AUC of **0.84**.
- Ensemble Voting Classifier performed slightly better with PR AUC of **0.35** and ROC AUC of **0.84**.
- **AGE**, **GENERAL HEALTH CONDITION**, and **POVERTY LEVEL** were consistently among the most important factors

Next Steps

- Try predicting on "Likely Has Had Psychological Distress in the Past Month". Much more imbalanced but likely more useful
- Use imbalanced-learn library to try variety of balancing algorithms
- Try using all the columns, even those not answered or applicable, impute missing values
- Try regression to predict on K6 scale, not just binary classification of 13 and above
- Could do more manual manipulation of ordinal categorical variables